CSIOVDB: a microarray gene expression database of epithelial ovarian cancer subtype

Tuan Zea Tan[1], He Yang[1], Jieru Ye[1], Jeffrey Low[2], Mahesh Choolani[2], David Shao Peng Tan[1,2,3], Jean-Paul Thiery[1,4,5], Ruby Yun-Ju Huang[1,2,5,6]

[1] Cancer Science Institute of Singapore, National University of Singapore, Center for Translational Medicine, Singapore 117599

[2] Department of Obstetrics and Gynecology, National University Health System, Singapore 119228

[3] Department of Haematology-Oncology, National University Hospital, Singapore 119074

[4] Institute of Molecular and Cell Biology, A*STAR, Proteos, Singapore 138673

[5] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117596

[6] Department of Anatomy, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117596

Correspondence to: Ruby Yun-Ju Huang, **e-mail:** ruby_yj_huang@nuhs.edu.sg
Jean-Paul Thiery, **e-mail:** bchtjp@nus.edu.sg

Keywords: *ovarian cancer, microarray gene expression, molecular subtype, database*

Received: *July 16, 2015* ***Accepted:*** *October 26, 2015* ***Published:*** *November 07, 2015*

ABSTRACT

Databases pertaining to various diseases provide valuable resources on particular genes of interest but lack the molecular subtype and epithelial-mesenchymal transition status. CSIOVDB is a transcriptomic microarray database of 3,431 human ovarian cancers, including carcinoma of the ovary, fallopian tube, and peritoneum, and metastasis to the ovary. The database also comprises stroma and ovarian surface epithelium from normal ovary tissue, as well as over 400 early-stage ovarian cancers. This unique database presents the molecular subtype and epithelial-mesenchymal transition status for each ovarian cancer sample, with major ovarian cancer histologies (clear cell, endometrioid, mucinous, low-grade serous, serous) represented. Clinico-pathological parameters available include tumor grade, surgical debulking status, clinical response and age. The database has 1,868 and 1,516 samples with information pertaining to overall and disease-free survival rates, respectively. The database also provides integration with the copy number, DNA methylation and mutation data from TCGA. CSIOVDB seeks to provide a resource for biomarker and therapeutic target exploration for ovarian cancer research.

INTRODUCTION

Every year, it is estimated that 238,700 women will develop ovarian cancer worldwide. Epithelial ovarian cancer (EOC) is the fifth-most common cause of female cancer death, with an estimated 151,900 deaths [1]. Even though the 5-year survival rates for localized, regional, and distant ovarian cancer are 91%, 72% and 27%, respectively, 61% of cases are presented at a stage when the disease is already widely metastatic [2], explaining the high mortality rate for this disease [3]. The majority of EOC patients respond well to first line platinum-based chemotherapy but about two-thirds of patients will eventually relapse with disease regardless of the initial clinical response [3]. Patients with recurrent EOC may initially respond to further chemotherapeutic agents but eventually develop chemoresistant disease and succumb to their illness.

At least 5 different histological subtypes of EOC exist and may reflect the clinical heterogeneity of this disease [4] in terms of chemotherapeutic response and outcome. Recently, it is becoming increasingly clear from the molecular analysis of EOC that this is also a molecularly heterogeneous disease [5–7]. While the relative clinical significance of these individual EOC molecular subtypes, as defined by high-throughput transcriptomics, remains unclear, recent data suggest that the gene expression profiles of EOC may have predictive value in determining patient benefit from targeted therapeutic agents such as bevacizumab (Avastin®) in frontline therapy [8]. Specifically, the mesenchymal/C1 [5–7] and Stem-A/Proliferative/C5 [5–7] subtypes were demonstrated to respond better to bevacizumab-containing regimen in ICON7 trials, with improvements in progression-free survival of 8.1 and 10.1 months, respectively [9]. Likewise, our group has also previously described that the Stem-A molecular subtype of EOC is sensitive to microtubule-targeted compounds such as vincristine and vinorelbine. On the other hand, the epithelial-mesenchymal transition (EMT) spectrum based on gene expression has also been described in EOC, where differential responses for epithelial-like and mesenchymal-like ovarian cancers have been reported; for example, mesenchymal ovarian cancer is reported to be more sensitive to cisplatin and benefits from a paclitaxel-containing treatment regimen [10, 11]. These studies suggest the clinical relevance of ovarian cancer molecular subtyping, and the potential to identify targeted therapies utilizing the molecular subtyping or EMT status.

Databases such as cBioportal [12, 13] and KMplotter [14] offer valuable resources to investigate genes of interest in various diseases, including ovarian cancer. In addition, OvMark [15], a database dedicated to investigate mRNA and miRNA expression in ovarian cancer provided tremendous insight into progression of the disease. However, none provide an indication of molecular subtype or EMT status. In this work, we built a transcriptomics database of human ovarian cancer, referred to as CSIOVDB (Ovarian cancer database of Cancer Science Institute Singapore; http://csibio.nus.edu.sg/CSIOVDB/CSIOVDB.html), which is furnished with molecular subtype and EMT information. Through this database, we seek to provide a complementary resource for gene expression profiling in ovarian cancer, particularly, the differential expression of molecular subtypes and its correlation with the EMT status. By delineating these transcriptomic subtypes and EMT status of EOC, it is envisaged that our database will facilitate further strategies to explore and guide targeted therapeutic approaches in this challenging disease.

RESULTS

Profiles of CSIOVDB

CSIOVDB comprises 3,431 microarray samples from 48 cohorts of private, in-house and public human ovarian cancer datasets (Materials and Methods; Figure 1; Suppl. Figure 1; Suppl. Table 1). The database contains 3,261 unique samples of mainly primary and metastatic

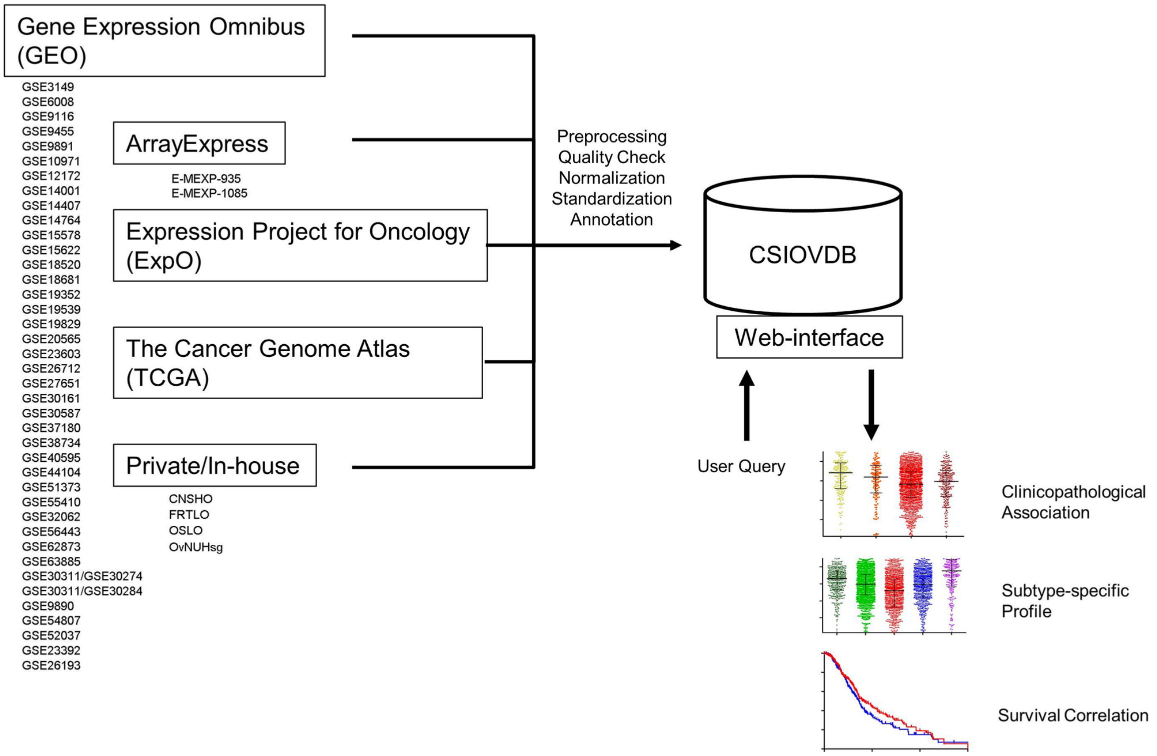


Figure 1: Structure of CSIOVDB. CSIOVDB housed data of human ovarian carcinoma from GEO, ArrayExpress, TCGA, ExpO, and private/in-house. The data from different cohorts were compiled, and subjected to quality check, RMA normalization, and standardization to remove batch effect. Clinical annotation was extracted from the data repository or original publication. User can query a gene of interest to CSIOVDB using a browser. Output of the query includes clinical association of the gene expression, gene expression profiles in different histologies, molecular subtypes, as well as survival correlation. Molecular subtype-specific correlation is also provided.

ovarian cancers (91.49%), as well as fallopian tube carcinoma (0.44%), peritoneal carcinoma (1.45%), metastasis to the ovary from elsewhere (1.95%), and ovarian cancer stroma (1.065). Non-cancerous samples constitute 3.63% of CSIOVDB and include normal ovarian surface epithelium (2.66%), normal ovary stroma (0.24%), and normal fallopian tube (0.73%; Figure 2A). Epithelial ovarian cancer is the main component of CSIOVDB; non-epithelial ovarian cancer, such as ovarian germ cell tumors, sex-cord stromal, and sarcoma (Figure 2B) comprise less than 1% of the database. Note, however, that CSIOVDB does not mirror the actual frequency of ovarian cancer, where non-epithelial ovarian cancer accounts for 10% of all ovarian cancers [16]. In terms of morphology, high-grade serous ovarian cancer is the most prevalent (73.75%) in CSIOVDB; this closely follows the 70% prevalence of high-grade serous ovarian cancer [4]. Ovarian cancers of other histologies, however, are slightly under-represented in CSIOVDB (CSIOVDB% vs reported%): mucinous (2.36% vs 3%), endometrioid (5.61% vs 10%), clear cell (4.43% vs 10%), and serous with low malignant potential (3.21% vs 5%) [4]. Late- (III & IV), and early-stage (I & II) tumor samples represent 65.35% and 12.72%, respectively, of the database (Figure 2C). The median age of the CSIOVDB samples is 58 years (Figure 2D). No menopausal information is available. Ovarian cancer grading is assessed either by FIGO (64.4%) or by the University of Texas M. D. Anderson Cancer Center [17] system (1.6%). High-grade ovarian cancers form the majority of CSIOVDB (63.27%; Figure 2E). Optimal (27.35%) and suboptimal (15.19%) surgical debulking status is also noted (Figure 2F), as this status is associated with ovarian cancer survival [18]. Surprisingly, whereas the surgical debulking status is associated with survival, this parameter does not contribute significantly to the molecular differences in ovarian cancer (Suppl. Figure 6D). Overall and disease-free survival data are available for 1,868 and 1,516 samples, respectively, with a median overall survival of 31.67 months and median disease-free survival of 17.09 months (Suppl. Table 2). Finally, molecular subtyping and EMT scores are provided in CSIOVDB. The database comprises 11.75% of ovarian cancer with an Epi-A subtype, 29.04% with Epi-B, 29.01% with Mes, 19.2% with Stem-A and 8.23% with Stem-B ovarian cancer; this spread of tumors mirrors the distribution of previous analyses [5] (Suppl. Figure 6A). Thus, overall, CSIOVDB represents a large and diverse collection of ovarian cancer that could be useful for assessing a gene of interest.

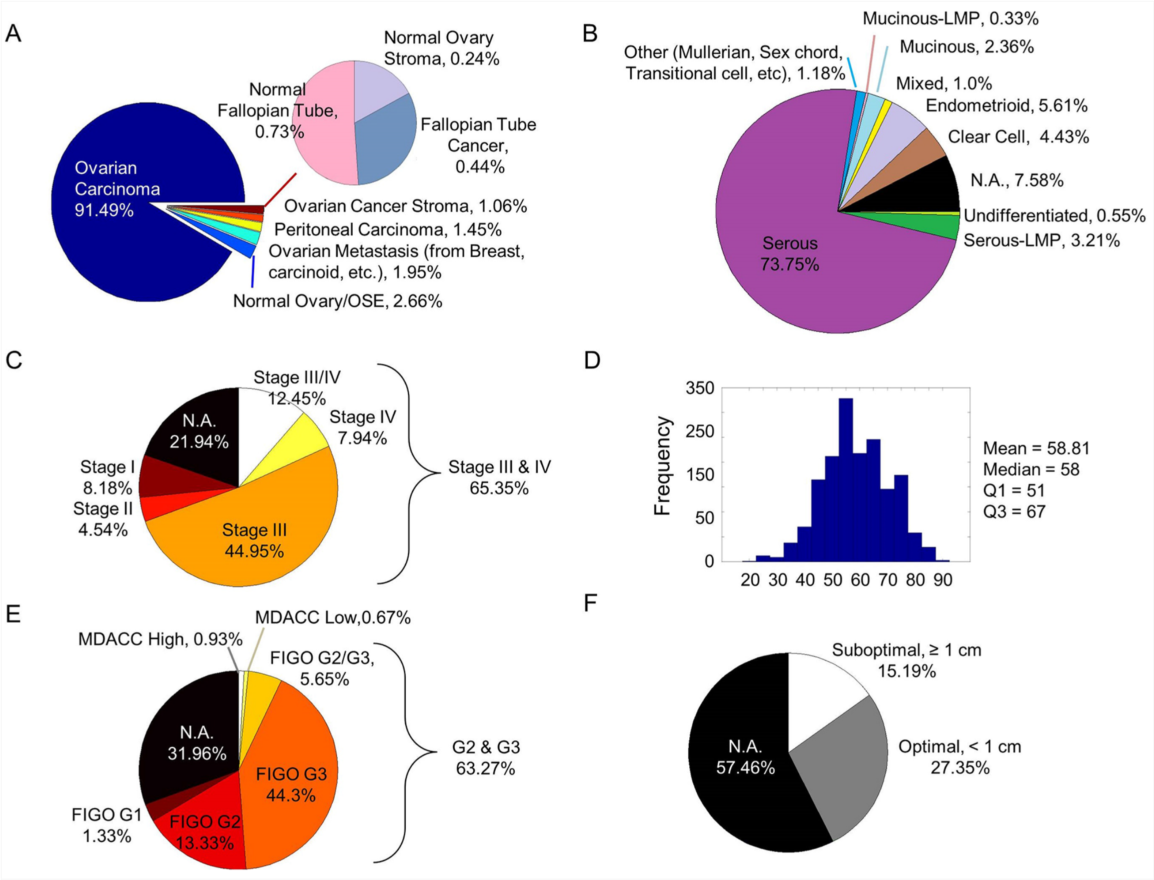


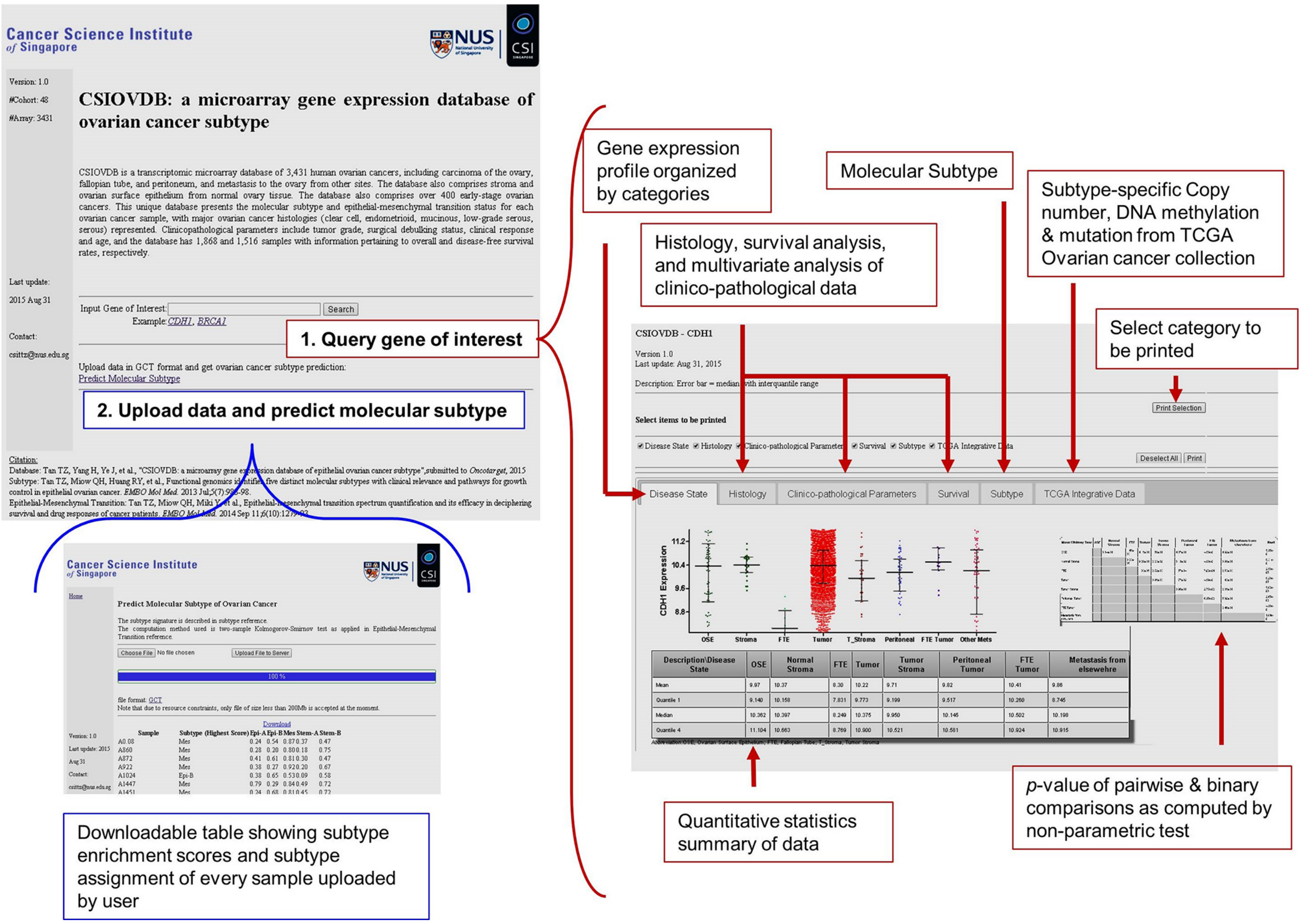
Figure 2: Clinico-pathological profiles of CSIOVDB. Pie charts or histogram showing distribution of various parameters in CSIOVDB: **A.** disease state, **B.** WHO histology, **C.** FIGO staging, **D.** age, **E.** FIGO grading or MDACC two-tier grading, and **F.** surgery debulking status. Abbreviation: N.A., not available.

Features of CSIOVDB

A screenshot of CSIOVDB is given in Figure 3 summarizing the features available. At the main page, there are two functions available: first, users can choose to query a gene of interest to CSIOVDB. At the result page of gene queried, the expression profiles of gene of interest are organized into different categories: disease state, histology, clinico-pathological parameters, and molecular subtype. In addition, molecular subtype-specific copy number, mutation, and DNA methylation profile from TCGA of the queried gene are provided in a separate tab. Quantitative statistics such as mean, median, upper and lower quantiles are available. Pairwise and binary significance assessments were performed using Mann-Whitney, Spearman correlation coefficient, or log-rank test. Multivariate Cox regression of queried gene and clinico-pathological parameters was also performed. User can select a subset of categories to be printed. Second, users can upload dataset for computation of ovarian molecular subtype. The computation method is based on two-sample Kolmogorov-Smirnov test and a subtype signature as described in previous work [5]. Computation of EMT score can be requested through email or from [10].

Querying a gene of interest

CSIOVDB seeks to provide users with the expression profiles of certain genes of interest relevant to ovarian cancer; in particular, the molecular subtype distribution and subtype-specific outcomes in terms of overall survival and disease-free (progression- and recurrence-free) survival. Figure 4 and Table 1 show a subset of outputs available from CSIOVDB when the gene *CDH1* is queried (http://csibio.nus.edu.sg/CSIOVDB/pages/CSIOVDB_CDH1.html). *CDH1* is an epithelial marker that codes for E-cadherin. The loss of E-cadherin expression has been linked to cancer progression and metastasis [19] and shown to display a differential expression profile in ovarian cancer (Figure 4). Interestingly, there is no significant difference in *CDH1* expression between ovarian cancer and normal ovarian surface epithelium (p = 0.415; Figure 4A), and this adheres to the previous findings that ovarian surface



Cancer Science Institute of Singapore

Version: 1.0

#Cohort: 48

#Array: 3431

CSIOVDB: a microarray gene expression database of ovarian cancer subtype

CSIOVDB is a transcriptomic microarray database of 3,431 human ovarian cancers, including carcinoma of the ovary, fallopian tube, and peritoneum, and metastasis to the ovary from other sites. The database also comprises stroma and ovarian surface epithelium from normal ovary tissue. The database also comprises over 400 early-stage ovarian cancers. This unique database presents the molecular subtype and epithelial-mesenchymal transition status for each ovarian cancer sample, with major ovarian cancer histologies (clear cell, endometrioid, mucinous, low-grade serous, serous) represented. Clinicopathological parameters include tumor grade, surgical debulking status, clinical response and age, and the database has 1,868 and 1,516 samples with information pertaining to overall and disease-free survival rates, respectively.

Last update: 2015 Aug 31

Contact: csittz@nus.edu.sg

Input Gene of Interest: Search

Example: *CDH1*, *BRCA1*

1. Query gene of interest

Upload data in GCT format and get ovarian cancer subtype prediction:
Predict Molecular Subtype

2. Upload data and predict molecular subtype

Citation:
Database: Tan TZ, Yang H, Ye J, et al., "CSIOVDB: a microarray gene expression database of epithelial ovarian cancer subtype",submitted to *Oncotarget*, 2015
Subtype: Tan TZ, Miow QH, Huang RY, et al., Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. *EMBO Mol Med.* 2013 Jul;5(7):983-98.
Epithelial-Mesenchymal Transition: Tan TZ, Miow QH, Miki Y, et al., Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med.* 2014 Sep 11;6(10):1279-93.

Cancer Science Institute of Singapore

Home

Predict Molecular Subtype of Ovarian Cancer

The subtype signature is described in subtype reference.
The computation method used is two-sample Kolmogorov-Smirnov test as applied in Epithelial-Mesenchymal Transition reference.

Choose File No file chosen | Upload File to Server

100 %

file format: GCT
Note that due to resource constraints, only file of size less than 200Mb is accepted at the moment.

Download

Sample	Subtype (Highest Score)	Epi-A	Epi-B	Mes	Stem-A	Stem-B
A0.08	Mes	0.24	0.54	0.87	0.37	0.47
A860	Mes	0.28	0.20	0.80	0.18	0.75
A872	Mes	0.41	0.61	0.81	0.30	0.47
A922	Mes	0.38	0.27	0.92	0.20	0.67
A1024	Epi-B	0.38	0.65	0.53	0.09	0.58
A1447	Mes	0.79	0.29	0.84	0.49	0.72
A1451	Mes	0.24	0.68	0.81	0.45	0.72

Version: 1.0

Last update: 2015 Aug 31

Contact: csittz@nus.edu.sg

Downloadable table showing subtype enrichment scores and subtype assignment of every sample uploaded by user

Gene expression profile organized by categories

Molecular Subtype

Subtype-specific Copy number, DNA methylation & mutation from TCGA Ovarian cancer collection

Histology, survival analysis, and multivariate analysis of clinico-pathological data

Select category to be printed

CSIOVDB - CDH1

Version 1.0
Last update: Aug 31, 2015

Description: Error bar = median with interquantile range

Select items to be printed

☑ Disease State ☑ Histology ☑ Clinico-pathological Parameters ☑ Survival ☑ Subtype ☑ TCGA Integrative Data

Description\Disease State	OSE	Normal Stroma	FTE	Tumor	Tumor Stroma	Peritoneal Tumor	FTE Tumor	Metastasis from elsewehre
Mean	9.97	10.37	8.30	10.22	9.71	9.82	10.41	9.86
Quantile 1	9.140	10.158	7.831	9.773	9.199	9.517	10.260	8.745
Median	10.362	10.397	8.249	10.375	9.950	10.146	10.502	10.198
Quantile 4	11.104	10.663	8.769	10.900	10.521	10.581	10.924	10.915

Abbreviation:OSE, Ovarian Surface Epithelium; FTE, Fallopian Tube; T_Stroma, Tumor Stroma

Quantitative statistics summary of data

p-value of pairwise & binary comparisons as computed by non-parametric test

Figure 3: Screenshot of CSIOVDB. Snapshot of CSIOVDB showing functions available on CSIOVDB. Firstly, query a gene of interest; and secondly, predict ovarian cancer subtype. Red or blue boxes indicate the features available on the main page: queried gene page and the page of predicting ovarian cancer subtype.

epithelium has both mesenchymal and epithelial features [20]. Neither peritoneal (p = 0.0177) nor fallopian tube carcinoma (p = 0.459) shows differential expression of *CDH1* with ovarian carcinoma. As expected, ovarian cancer stroma, which is more mesenchymal-like, has a significantly lower *CDH1* level than its carcinoma counterpart (p = 0.004). From a histological perspective, high-grade serous and endometrioid ovarian cancers have the lowest *CDH1* expression (Figure 4B). The less-aggressive serous carcinoma with low malignant potential has significantly higher *CDH1* expression compared with high-grade serous ovarian cancer (p = 4.12E-08). Also not surprisingly, the more metastatic and aggressive late-stage (p = 1.44E-12) and high-grade (p = 4.42E-05) ovarian cancer have significantly lower *CDH1* expression (Figure 4C); however, there is no difference in *CDH1* expression for clinical response (Figure 4C). Since *CDH1* is an epithelial marker, it displays a negative correlation with EMT score (*Rho* = -0.32; Figure 4D). No correlation was observed between *CDH1* expression and age (*Rho* = -0.067; Figure 4D). Importantly, *CDH1* was found to be lowest in the Mes subtype (p = 7.15E-38; Figure 4E), which is enriched with metastatic ovarian cancers [5] and supports the conjecture that a loss of *CDH1* promotes metastasis [19]. Despite evidence for the loss of *CDH1* in tumors with a Mes subtype, as well as the association between a Mes subtype and metastasis, our database shows no correlation for *CDH1* with overall survival for any of the subtypes (Figure 4E, 4F). This is likely due to the fact that *CDH1* is expressed in Stem-A, another poor survival subtype (Figure 4E, Suppl. Figure 6B).

This example of the data obtained as an output from CSIOVDB also demonstrates the range of descriptive statistics provided (mean, median, quantiles and statistical significance evaluations). Furthermore, a multivariate Cox regression analysis of *CDH1* gene expression levels and known ovarian cancer prognostic factors, such as stage, grade, surgical debulking status, histology and age, is provided (Table 1). Importantly, CSIOVDB provided not only gene expression profiles of molecular subtypes, but also subtype-specific survival outcomes.

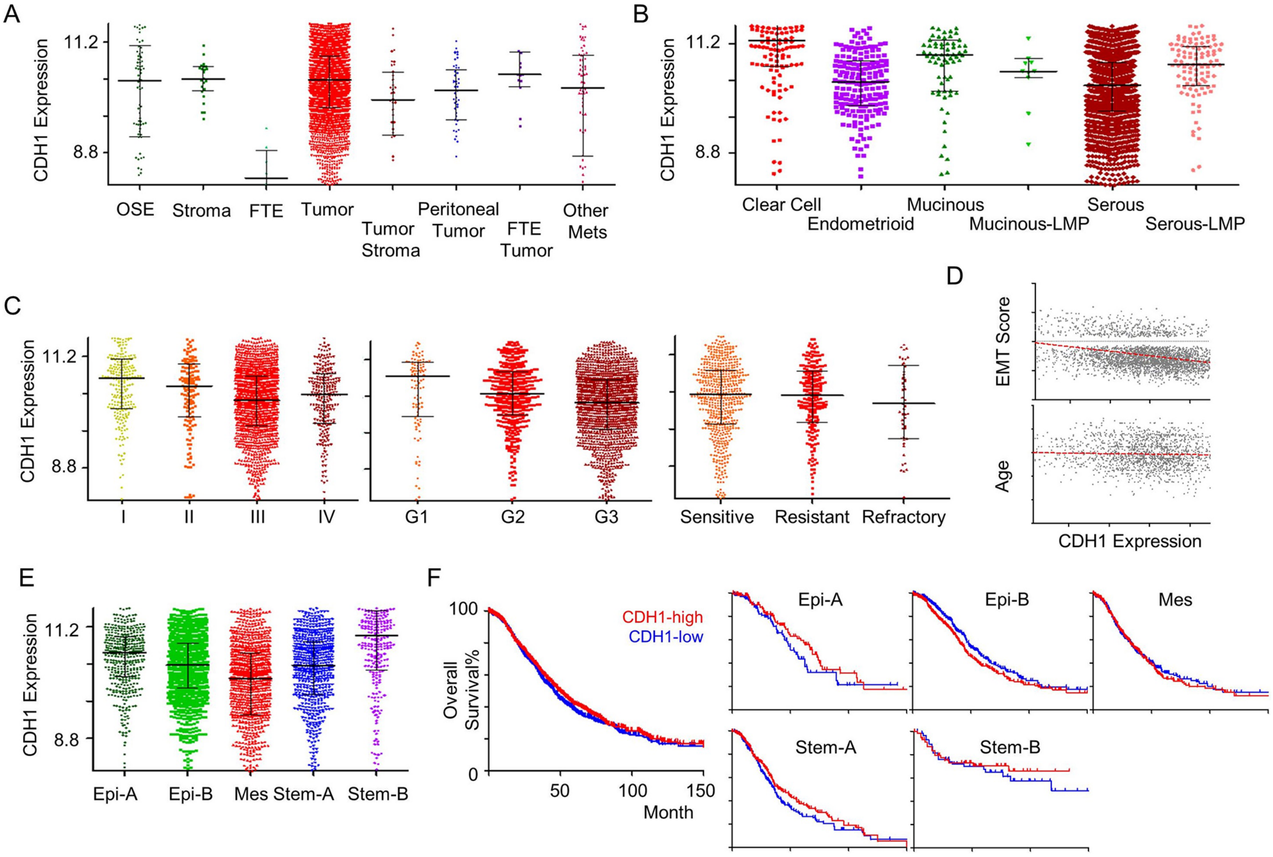


Figure 4: *CDH1* gene expression in ovarian cancer. An example of CSIOVDB outputs for a gene queried. Gene expression profiles of *CDH1* in ovarian cancer disease state **A.** histology **B.** FIGO stage **C.** Left panel FIGO grade C. Middle panel. clinical response C. Right panel epithelial-mesenchymal transition (EMT) score **D.** Upper panel age D. Lower panel molecular subtype **E.** and overall survival **F.** Left panel overall survival within subtype **F.** Right panel. Median expression was used to define CDH1-high and CDH1-low groups. Linear regression line fit is shown in red in (D). Abbreviation: OSE, ovarian surface epithelium; FTE, fallopian tube epithelium; Mets, metastasis; LMP, low-malignant potential; Epi, epithelial; Mes, mesenchymal; Stem, stem-like. Error bar is median ± quantile.

Table 1: Multivariate Cox regression analysis of ovarian cancer overall survival

Prognostic Factor	Cox Coefficient	*p*-value
Stage (I,II vs III, IV)	1.40	1.0E-5
Grade (G1 vs G2,G3)	0.65	0.0931
Surgical Debulking (optimal vs suboptimal)	0.2	0.0373
Histology (Non-serous vs serous)	1.29	0.0105
Age (< 55 vs ≥ 55)	0.16	0.0777
CDH1	-0.05	0.5966

DISCUSSION

CSIOVDB is a transcriptomics database of human ovarian cancer comprising 3,431 microarray data from 48 cohorts. Each sample is coupled with histology, molecular subtype and EMT information to enable the user to explore and investigate genes of interest. To the best of our knowledge, this is the first database that integrates molecular subtype and EMT with ovarian cancer. This unique feature of CSIOVDB allows interrogation of subtype-specific expression as well as survival profiles. It is our hope that the database will provide a complementary resource to existing general [12–14] or databases specifically dedicated for ovarian cancer [15] for the investigation of clinical associations for genes of interest and, more importantly, for the localization and assessment of potential biomarkers or therapeutic targets for ovarian cancer. The previously determined preferential responses of mesenchymal-like ovarian cancer to paclitaxel [10], and platinum [11], the chemosensitivity to vincristine and vinorelbine in Stem-A ovarian cancer [5], and the chemosensitivity to bevacizumab in mesenchymal and Stem-A (proliferative) ovarian cancer [9] all suggest the feasibility of targeted therapeutics for ovarian cancer and, by extension, the utility of CSIOVDB.

However, it is important to note that ovarian cancer is an extremely heterogeneous disease [21]—95% of ovarian cancers are clonally heterogeneous and many have four or more subclones [22]. Not surprisingly, therefore, most ovarian cancer exhibit properties of multiple subtypes [23]. There are as many as 82% of the TCGA and 42% of the Mayo ovarian cancer cohorts displayed properties of at least two subtypes [24]. Thus, ovarian cancer treatment regimens may require a multi-agent approach, targeting the different subclones that exhibit diverse subtypes; it is plausible that targeting one subclone will only allow another to take over [25]. Adding to this complexity, studies have shown that the molecular subtype of a tumor may change post-chemotherapy [26], prompting the need for a continuous subtype re-assessment during the course of chemotherapy. In CSIOVDB, samples are assigned to the strongest phenotypic subtype (as a subtype gene expression signature [5]). Given that each subtype is represented by a sufficiently large number of samples, the influence of sample heterogeneity on subtype gene expression may be mitigated.

Aside from heterogeneity within a sample, a further caveat of using CSIOVDB is that the database is built based on public data from different repositories contributed by various authors and laboratories. While we have made every endeavor to limit the effect of a center-related batch effect, factors [27] such as reagents, protocols, procedures, elapsed time from sample collection, therapy regimen, and many others, could not be accounted for because of a lack of data. Thus, the data obtained through CSIOVDB should be viewed as a preliminary analysis, and users are urged to exercise due diligence in validating their findings.

On a side note, to ensure the relevance of CSIOVDB, we plan to annually update CSIOVDB to include new publicly available data as well as update of clinical and annotation data. In addition, we intend to replicate the process of building CSIOVDB to other cancers and allow users to query a gene of interest in multiple cancers and molecular subtypes.

MATERIALS AND METHODS

Eligibility criteria

As our purpose was to compile a database of broader generalizations and larger sample size, we adopted less stringent eligibility criteria [28]. Three private, one in-house and 44 publicly available microarray ovarian cancer datasets from Gene Omnibus (GEO; http://www.ncbi.nlm.nih.gov/gds), ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), Expression Project for Oncology (ExpO; http://www.intgen.org/), and The Cancel Genome Atlas (TCGA; http://cancergenome.nih.gov/) were downloaded by Jan 2015 (Figure 1; Suppl. Table 1). Only datasets obtained using Affymetrix Microarrays HG-U133A (16.38%), HG-U133A2 (2.88%), HG-HT-U133A (17.25%), HG-U133-Plus2 (49.2%), and human gene 1.0 ST (14.29%) were used. These datasets are inclusive of primary and metastatic ovarian cancers, fallopian tube carcinoma, peritoneal carcinoma, ovarian cancer stroma, and normal ovarian surface epithelium, fallopian tube, and stroma tissues (Figure 2). No limit was imposed on the race, pre-treatment history or medical condition, stage, grade, or histology of the disease.

National University Hospital cohort

Frozen archival EOC samples from Department of Obstetrics & Gynecology, National University Hospital of Singapore were collected from 2006 to 2014. Frozen tumor samples were kept frozen at all times prior to evaluation. Each frozen tumor sample was pounded to a fine powder in liquid nitrogen using a pre-chilled mortar and pestle, and the powdered sample collected into a pre-chilled microfuge tube. Samples were homogenized in Trizol (Life Technologies, Carlsbad, CA) using a sterile 1-ml syringe and a 21-G hypodermic needle (BD Precision, Oxford, AL). After homogenization, RNA was purified using a Qiagen miRNeasy kit, as per manufacturer's protocol (Hilden, Germany). RNA sample quality was determined by Eukaryote Total RNA Nano Series II, 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). RNA samples with a RIN value above 6.5 were used for the Affymetrix GeneChip® Human Gene 1.0 ST Array (Affymetrix, Inc., Santa Clara, CA). Data has been deposited in GEO with the accession GSE69207.

Clinico-pathological parameters

All clinical and pathological information were extracted either from ArrayExpress and GEO, or from publications associated with the data. Samples without sufficient information are flagged as 'Not Available'. In some categories, abstraction or simplification was performed to have a sizeable group; for example, the histopathology of ovarian carcinoma, as classified according to the World Health Organization (WHO; http://www.who.int), is restricted to (high-grade) serous, borderline/low-malignant potential (LMP) serous, mucinous, endometrioid, and clear cell. Other less-prevalent ovarian cancers, such as Brenner, Signet ring cell, sex cord-gonadal stromal, and mixed Mullerian tumors, and ovarian cancers with mixed histologies are grouped as 'other' and 'mixed', respectively; no distinction is made for the International Federation of Gynecology and Obstetrics (FIGO; http://www.figo.org) group within each stage (e.g. stages IA, IB, IC are grouped as stage I). Grading of ovarian cancer was conducted using FIGO or the two-tier grading system proposed by the University of Texas M. D. Anderson Cancer Center [17]; these two grading systems usually show good correlation [17]. The optimal surgical debulking status is defined as having residual tumor of less than 1 cm. Clinical response is defined by either response evaluation criteria in solid tumors (RECIST) version 1.0 and above, or the serum level of CA-125; these two tests also show comparable results [29]. Pathological response is also available in some of the ovarian cancer samples, where pathological complete response is defined as no residual carcinoma or no residual invasive tumor. For simplicity, we categorized clinical response into 'sensitive' (RECIST complete response; pathological complete responder), 'resistant' (RECIST partial response, stable disease; pathological non-complete responder), and 'refractory' (RECIST progressive disease; pathological non-responder). Overall survival was computed by the difference between the date of last follow-up (or date of death) and the date of diagnosis, regardless of the cause. Disease-free survival encompasses progression-, local and distant recurrence-free survival.

Preprocessing of affymetrix expression data

A total of 3,431 arrays corresponding to 3,261 unique patients were collected. Prior to normalization, quality control was performed on the Affymetrix chips using R version 3.1.2 (2014–10-31) and Bioconductor packages (affy version 1.42.3, affyQCReport version 1.42) for 3′IVT arrays (HG-U133 series), or Affymetrix Power Tools version 1.15.2 for human exon array (human gene 1.0 ST). Details of R session information can be found in the Supplementary Information. Quality metrics and the following criteria were analyzed: average perfect-match (PM) intensity, background, scale factor, GAPDH 3′:5′ ratio, β-actin 3′:5′ ratio, area under the curve (AUC) of positive versus negative controls, relative log expression median, relative log expression inter-quantile range (Suppl. Figure 2). All chips passed at least one of the criteria, and hence, none of the samples was discarded.

For the post-quality check, the data from the 48 cohorts was combined and normalized using frozen robust multichip average (fRMA) [30] version 1.16. Annotation details for each platform, as required by fRMA, are given in the R session information in the Supplementary Information. Three datasets—U133A-U133P2-Gene1.0ST, U133P2-Gene1.0ST, and Gene1.0ST—were compiled to obtain probes/genes unique to each platform (Suppl. Figure 3). Probe matching was performed based on Affymetrix HG-U133-Plus2 to human gene 1.0 ST good- and perfect-matched probe-sets (the product sheet is available at http://www.affymetrix.com/support/technical/byproduct.affx?product=hugene-1_0-st-v1). In the case where multiple-to-one or one-to-multiple probes matched, only those probe-sets with highest intensities were kept. Annotation of the probes is based on the Affymetrix annotation version na34. The combined and normalized data were subsequently standardized using ComBat [31] to remove any batch effect by the cohorts and the centers. The influence of batch effect was assessed for clinico-pathological parameters and potential confounding factors: cohort, processing batch, centers and platform, pre- and post-ComBat standardization (Suppl. Figure 4A). A Kruskal-Wallis test of the first 45 principal components (>90% variance) and the parameters/factors indicates that the batch effect due to cohort, processing batch, center and platform were minimized post-standardization without removing differences due to clinico-pathological parameters (Suppl. Figure 4A; Suppl. Info). Inter-sample correlations and subtype concordance analyses of JPKO (GSE30311) samples available on both Affymetrix U133-Plus2 and human gene 1.0 ST provide additional support that the batch effect was not overwhelming in CSIOVDB (Suppl. Figures 4B & 4C; Suppl. Info).

Ovarian cancer molecular subtype and epithelial-mesenchymal transition (EMT) status

Gene expression values of ovarian cancer molecular subtype signatures [5] were extracted from the standardized dataset and subjected to clustering using Bioconductor ConcensusClusterPlus version 1.18. Parameters chosen were hierarchical clustering with agglomerative average linkage, with Euclidean distance and a sub-sampling ratio of 0.8 for 1000 iterations. The condition K_{max} was set to 5 to assign each sample to one of the ovarian cancer molecular subtypes: Epithelial-A (Epi-A), Epi-B, Mesenchymal (Mes), Stem-like (Stem)-A, or Stem-B (Suppl. Info). EMT scores for each sample were computed using an ovarian cancer-specific EMT signature and two-sample Kolmogorov-Smirnov test, as described previously [10].

TCGA ovarian cancer data integration

Level-3 and level-2 data of ovarian cancer SNP array, Illumina human 27K DNA methylation array and exome-sequencing were respectively downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) on August 31, 2015. Samples were matched by TCGA patient ID and samples with molecular subtype available (computed in this study) were used for analyses.

Multivariate cox regression analysis

Multivariate survival analyses of known ovarian cancer prognostic factors (stage, grade, age, histology and surgical debulking status) and gene expression were computed using Cox regression on 987 (overall survival) or 778 (disease-free survival) ovarian cancer samples with all clinico-pathological data available. In the multivariate Cox regression analysis, the factors and gene expression parameters are converted to binary states: stage is categorized as early (I, II) or late (III, IV); grade is categorized as low (G1) or high (G2, G3); age is categorized as young (< 55) or old (≥ 55); histology is categorized as non-serous or serous; and gene expression is categorized as low (< median expression) or high (≥ median expression).

Statistical analysis

Statistical analyses were conducted using Matlab ® R2012a version 7.14.0.739, and statistics toolbox version 8.0 (MathWorks; Natick, MA). Statistical significance of differential expression was evaluated using either Kruskal-Wallis or Mann-Whitney *U*-test. A Spearman correlation coefficient test was applied to assess significance of correlation. Kaplan-Meier analyses were conducted using GraphPad Prism ® version 5.04 (GraphPad Software; La Jolla, CA). Statistical significance of the Kaplan-Meier analysis was calculated by log-rank test.

ACKNOWLEDGMENTS

We thank Dr. R. Jackson for her careful English editing.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

GRANT SUPPORT

This research is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative to Cancer Science Institute of Singapore (JPT and HY); National University of Singapore Start-up Grant (JPT); National University Cancer Institute, Singapore Centre Grant (RYH).

.

Author contributions

JPT and RYH devised the project and obtained funding; TZT, RYH, HY, JPT wrote the paper; TZT and HY performed bioinformatics analyses.

REFERENCES

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J and Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians. 2015; 65:87–108.
2. Siegel R, Ma J, Zou Z and Jemal A. Cancer statistics, 2014. CA: a cancer journal for clinicians. 2014; 64:9–29.
3. Jayson GC, Kohn EC, Kitchener HC and Ledermann JA. Ovarian cancer. Lancet. 2014; 384:1376–1388.
4. Prat J. Ovarian carcinomas: five distinct diseases with different origins, genetic alterations, and clinicopathological features. Virchows Archiv : an international journal of pathology. 2012; 460:237–249.

5. Tan TZ, Miow QH, Huang RY, Wong MK, Ye J, Lau JA, Wu MC, Bin Abdul Hadi LH, Soong R, Choolani M, Davidson B, Nesland JM, Wang LZ, Matsumura N, Mandai M, Konishi I, et al. Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. EMBO molecular medicine. 2013; 5:983–998.

6. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615.

7. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B, Traficante N, Fereday S, Hung JA, Chiew YE, Haviv I, Australian Ovarian Cancer Study G, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008; 14:5198–5208.

8. Gourley C, McCavigan A, Perren T, Paul J, Michie CO, Churchman M, WIlliams A, McCluggage WG, Parmar M, Kaplan RS, Hill LA, Halfpenny IA, O'Brien EJ, Raji O, Deharo S, Davison T, et al. Molecular subgroup of high-grade serous ovarian cancer (HGSOC) as a predictor of outcome following bevacisumab. Journal of Clinical Oncology. 2014; 32:5502.

9. Clinic M. (2014). Ovarian cancer subtypes may predict response to bevacizumab. ScienceDaily).

10. Tan TZ, Miow QH, Miki Y, Noda T, Mori S, Huang RY and Thiery JP. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO molecular medicine. 2014; 6:1279–1293.

11. Miow QH, Tan TZ, Ye J, Lau JA, Yokomizo T, Thiery JP and Mori S. Epithelial-mesenchymal status renders differential responses to cisplatin in ovarian cancer. Oncogene. 2015; 34:1899–1907.

12. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C and Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling. 2013; 6:pl1.

13. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C and Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012; 2:401–404.

14. Gyorffy B, Surowiak P, Budczies J and Lanczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. PloS one. 2013; 8:e82241.

15. Madden SF, Clarke C, Stordal B, Carey MS, Broaddus R, Gallagher WM, Crown J, Mills GB and Hennessy BT. OvMark: a user-friendly system for the identification of prognostic biomarkers in publically available ovarian cancer gene expression datasets. Molecular cancer. 2014; 13:241.

16. Smith HO, Berwick M, Verschraegen CF, Wiggins C, Lansing L, Muller CY and Qualls CR. Incidence and survival rates for female malignant germ cell tumors. Obstetrics and gynecology. 2006; 107:1075–1085.

17. Malpica A, Deavers MT, Lu K, Bodurka DC, Atkinson EN, Gershenson DM and Silva EG. Grading ovarian serous carcinoma using a two-tier system. The American journal of surgical pathology. 2004; 28:496–504.

18. du Bois A, Reuss A, Pujade-Lauraine E, Harter P, Ray-Coquard I and Pfisterer J. Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: a combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials: by the Arbeitsgemeinschaft Gynaekologische Onkologie Studiengruppe Ovarialkarzinom (AGO-OVAR) and the Groupe d'Investigateurs Nationaux Pour les Etudes des Cancers de l'Ovaire (GINECO). Cancer. 2009; 115:1234–1244.

19. Thiery JP. Epithelial-mesenchymal transitions in tumour progression. Nature reviews Cancer. 2002; 2:442–454.

20. Auersperg N, Wong AS, Choi KC, Kang SK and Leung PC. Ovarian surface epithelium: biology, endocrinology, and pathology. Endocrine reviews. 2001; 22:255–288.

21. Waldron L, Riester M and Birrer M. Molecular subtypes of high-grade serous ovarian cancer: the holy grail? Journal of the National Cancer Institute. 2014; 106.

22. Lohr JG, Stojanov P, Carter SL, Cruz-Gordillo P, Lawrence MS, Auclair D, Sougnez C, Knoechel B, Gould J, Saksena G, Cibulskis K, McKenna A, Chapman MA, Straussman R, Levy J, Perkins LM, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer cell. 2014; 25:91–101.

23. Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ, Fereday S, Lawrence M, Carter SL, Mermel CH, Kostic AD, Etemadmoghadam D, Saksena G, Cibulskis K, Duraisamy S, Levanon K, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. The Journal of clinical investigation. 2013; 123:517–525.

24. Konecny GE, Wang C, Hamidi H, Winterhoff B, Kalli KR, Dering J, Ginther C, Chen HW, Dowdy S, Cliby W, Gostout B, Podratz KC, Keeney G, Wang HJ, Hartmann LC, Slamon DJ, et al. Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. Journal of the National Cancer Institute. 2014; 106.

25. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013; 152:714–726.

26. Lim SK, Lee MH, Park IH, You JY, Nam BH, Kim BN, Ro J, Lee KS, Jung SY, Kwon YM and Lee ES. Impact of Molecular Subtype Conversion of Breast Cancers after

Neoadjuvant Chemotherapy on Clinical Outcome. Cancer research and treatment : official journal of Korean Cancer Association. 2015. doi: 10.4143/crt.2014.262. [Epub ahead of print]

27. de Gramont A, Watson S, Ellis LM, Rodon J, Tabernero J, de Gramont A and Hamilton SR. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. Nature reviews Clinical oncology. 2015; 12:197–212.

28. George SL. Reducing patient eligibility criteria in cancer clinical trials. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 1996; 14:1364–1370.

29. Rustin GJ, Timmers P, Nelstrop A, Shreeves G, Bentzen SM, Baron B, Piccart MJ, Bertelsen K, Stuart G, Cassidy J and Eisenhauer E. Comparison of CA-125 and standard definitions of progression of ovarian cancer in the intergroup trial of cisplatin and paclitaxel versus cisplatin and cyclophosphamide. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006; 24:45–51.

30. McCall MN, Bolstad BM and Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010; 11:242–253.

31. Johnson WE, Li C and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–127.